Pure Water Structure and Hydration Forces for Protein Folding

Teresa Head-Gordon*

Department of Bioengineering, University of California, Berkeley, CA 94720

Greg Hura

Graduate Group in Biophysics, University of California, Berkeley and Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Jon M. Sorenson

Department of Chemistry, University of California, Berkeley, Berkeley CA 94720

Robert M. Glaeser

Department of Molecular and Cell Biology, University of California, Berkeley, & Life Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Abstract

This paper summarizes results of our recent work on pure water scattering and models for aqueous hydration potentials of mean force for protein folding.

†**To whom correspondence should be addressed**

Introduction

Experimental and theoretical studies on the folding of small globular proteins have established that they fold by a two-state mechanism, finding that the resulting folding is very cooperative, similar to a first-order phase transition in bulk materials. The cooperativity of folding is fundamentally connected to a free energy landscape thought to be funneled due to the presence of sufficient energetic biases that ultimately win over the loss of conformational entropy of the unfolded chain[1,2]. The free energy landscape theory has now provided a theoretical framework within which to develop minimalist protein folding models or all-atom protein folding simulations that have stronger quantitative connections to experiment [3-8].

Quantitative comparisons to protein folding experiments must of course connect back to the development of a free energy surface describing both the protein and its aqueous environment. Typically experiments guide the development of the underlying potential energy functions, but the experimental repertoire for characterizing aspects of this free energy surface are primarily focused on either the protein itself or separately on the pure water fluid. Our group has been involved in developing a better connection between the two systems through the combined use of both neutron and x-ray scattering experiments, molecular dynamics simulations and theoretical analysis on aqueous solutions of biological molecules over a range of concentrations to deepen our understanding of hydration in protein folding[9-11].

The scattering intensity contains information about the solute-solute correlations in water through the radial distribution function between solute centers, $g_c(r)$. The solute-solute radial distribution function in aqueous solution, $g_c(r)$, can be equated with the Boltzmann factor, $g_c(r) = e^{-W(r)/k_bT}$, where $W(r)$ defines an averaged potential, or a "potential of mean force", between the two solutes separated by a distance r. The information content of the solution scattering experiments is the net correlations between solute pairs that takes into account the complicated solvent environment, which can be extracted with the aid of simulations and equated with the thermodynamics of amino acid association in water.

One outcome of this approach is the discovery of "reduced" descriptions of hydrophobic solvation, i.e. the extraction of hydrophobic potentials of mean force (pmf) from the experimental intensity using simulation. We have focused on solutions of a common hydrophobic amino acid solute, N-acetyl-leucine-amide (NALA) or –methylamide (NALMA)[9-11] as a function of its concentration in water to demonstrate this approach. Dilute solution scattering studies of this amino acid monomer constitutes a model of the solvation structure and free energy of hydrophobic solute association during early protein self-assembly events when the local concentration of amino acids is relatively dilute and residues are well-hydrated[9,10]. We also have reported x-ray solution scattering results on the behavior of the hydrophobic NALMA solute in water at high concentrations[10,11] as a model of later folding events when significant spatial domains of the protein are driving toward the formation of the hydrophobic core. We have used these experimentally derived pmfs as reduced solvation descriptions to define new energy surfaces on which we have performed global optimizations to predict protein structure in the recent CASP4 competition with significant success in the more difficult protein targets of this contest[14]. In the results we briefly review this published work for the concentrated NALMA solutions.

At the same time, the simulations of the experimental intensity observable that are needed to extract the solute-solute correlations for these solutions also gives a direct assessment of the quality of the protein and water force fields themselves. We have determined that while the underlying force fields appear to be qualitatively consistent with experiment (we believe), the quantitative agreement is sometimes poor depending on the combination of protein force field and water model used. Certainly some of the error resides in the water model itself, and the quality of a given water model can be judged against a battery of experiments for structure and dynamics. In the results we describe how the ambiguity in past structural experiments has led us perform a high quality synchrotron experiment[12] and corresponding theoretical analysis[13] of the structure of ambient neat water that has proved important for

vetting existing water potentials, new non-polarizable and polarizable water models, and emerging simulation methodologies such as Car-Parrinello molecular dynamics.

Results

X-ray solution scattering measurements at concentration ratios of NALMA solute to water of 1:25 through to 1:100 were obtained, as well as that for pure water, some of which are shown in Figure 1. The x-ray curve for the more dilute concentrations is dominated by the main x-ray diffraction peak of water at room temperature at Q~2.0Å^{-1}. However, at a concentration of 1:50, a new feature appears at Q~0.8Å^{-1}, and develops into a peak at 1:25, the maximum concentration studied. The new diffraction at Q~0.8Å^{-1} reflects the formation of a fluid, but ordered, phase, the amount of which depends upon the total solute concentration, but whose internal structure is not sensitive to solute concentration.

We have used molecular dynamics simulations to interpret this new experimental feature at Q~0.8Å^{-1}. While we cannot simulate the time progression involved in the formation of solute distributions seen experimentally, as this would require MD simulations over very long time scales in order to reach the final equilibrated distribution of solutes, considerations of the mechanisms of how these solute configurations are reached are not important for this experiment. What is important is determining the final configurations of solutes that reproduce the static experimental observable.

We have focused therefore on what we view is a rough but complete representation of the possible distributions of solutes seen experimentally. First, we considered a fully dispersed and hydrated configuration of NALMA molecules in water at concentrations of solute to water of 1:24 and 1:48. A second class of solute configuration involves the formation of small molecular aggregates of solutes that range from mono-dispersed to clusters containing roughly two to six NALMAs in the most concentrated solutions. Concentration ratios of solute to water considered were in the range 1:24 to 1:100. Finally, we

consider the case that all NALMAs are configured into one cluster, for concentrations of 1:24, 1:42, and 1:48.

Figure 2 shows a comparison of the simulated data with the experimental data at the highest concentration of 1:25. The concentration dependence seen experimentally (all data not shown) is best reproduced by configurations of NALMA in which the solutes are maximally dispersed or involving small molecular clusters on the order of two to six NALMAs. When considering the best single cluster data, the scattering predicted for the smallest single cluster is too sharply defined and slightly shifted to a smaller Q-value. This gets worse for the larger-sized cluster (which is simulated in a larger box and is therefore more dilute) where there is a significant rise in intensity and shift away from Q ~0.8Å^{-1} (data not shown).

Our experimental and simulation results over a range of hydrophobic amino acid concentrations imply aqueous potentials of mean force with two free energy minima such as that shown in Figure 3, and by our model indicate the type of aqueous free energy biases in the folding free energy landscape for protein self-assembly. The importance of this result is that even at late stages of folding when the local concentration of hydrophobic amino acids is very high, significant water is still present and stabilizing hydrophobic amino acids at lengthscales where they are separated by a water layer. To further strengthen the connection between our monomer systems and proteins, we have found that simple order-of-magnitude estimates for the change in configurational entropy arising from the collapse of a dense solution are the same as estimates of the conformational entropy change experienced by a polymer chain in the later stages of collapse.

Figure 4 shows the better quantitative agreement with experiment using the TIP-FQ water model at mole ratios of solute to water of 1:25 when it is compared to the SPC simulation (Figure 2). A better quantitative comparison between simulation and experiment gives us more confidence in the result that dispersed to small molecular aggregates are formed at these high concentrations of hydrophobic solutes in

water. Because these are the newest generation force fields, we might ask how well they reproduce something as basic as ambient pure water structure.

To answer this question we have recently performed a new x-ray diffraction study at the Advanced Light Source (ALS) in Berkeley of liquid water under ambient conditions that takes advantage of various state-of-the-art features of a modern day experiment[12]. Improvements include a well characterized polarization correction, a high-level Compton scattering correction, higher energy x-rays that permit the use of thinner samples and reduces the need for a multiple scattering correction, more accurate intensities using a modern CCD image plate detector, and careful attention to the evaluation of the atomic form factors for extraction of radial distribution functions. The error bars show that our data has accuracy exceeding the differences when compared to the scattering curves of past x-ray in Figure 5.

We have found that the true charge density of the water molecule in the condensed phase requires a modification of the isolated atom scattering factors commonly used in the extraction of radial distribution functions from neutron and x-ray scattering data[13] according to:

$$I(Q) = \sum_{ij} x_i x_j f_i(Q) f_j(Q) \frac{\sin Qr_{ij}}{Qr_{ij}} + \sum_{i \le j} x_i x_j f_i(Q) f_j(Q) h_{ij}(Q) \qquad (1)$$

We know that the Debye approximation, the assumption of superposition of the standard atomic scattering factors, $f(Q)$, performs inadequately for gas phase water. A simple modification is to scale the atomic scattering factors by the proper factor which gives a value of 1.86D for the dipole moment of gas-phase water, i.e. multiply $f_O(Q)$ by 1.11 and $f_H(Q)$ by 0.56. The result is that this simple adjustment greatly improves agreement at small Q, but at the sacrifice of agreement at large Q. The reason for this is that the large Q tails of the atomic scattering factors probe the density profile of the core electrons of the individual atoms. The core density would be expected to change much less upon chemical bonding, and in fact the Debye expression gives excellent agreement with the essentially exact result at large Q. This

suggests a modification of the atomic scattering factors which rescales them properly at low Q, but retains their values at large Q. Such a modification is the following:

$$f'(Q) = \left[1 + (\alpha - 1)exp\left(-Q^2 / 2\delta^2\right)\right] f(Q) \quad (2)$$

where $f'(Q)$ is the modified atomic scattering factor (MASF), $f(Q)$ is the atomic scattering factor for the isolated atom, α is a scaling factor giving the redistribution of charge, and δ is a parameter to be fit, representing the extent of valence-electron delocalization induced by chemical bonding. For gas phase water, we choose α to correspond to the gas phase dipole moment. The unknown parameter δ can be fit by requiring the Debye expression curve to agree with the best available *ab initio* CI result. A single parameter choice of δ=2.2Å^{-1} for both oxygen and hydrogen MASF's was found to give excellent agreement when Eq. 2 was used to reproduce gas phase water scattering.[13] The advantage of the MASF formalism lies in the firmer foundation it provides for extraction of the oxygen-oxygen (OO) and possibly oxygen-hydrogen (OH) correlations from the experimental scattering curves. With the proper scaling, they allow the correct weighting of OO and OH correlations, allowing one to extract $g_{OO}(r)$ and not only a molecular centers radial distribution function.

Using the MASF's and a fitting of the x-ray intensity based on a linear combination of a representative basis set of $g_{OO}(r)$'s, we have determined a $g_{OO}(r)$ for water consistent with our recent experimental data gathered at the ALS that is different than the $g_{OO}(r)$ reported by other x-ray and neutron scattering experiments. Compared to past experiments[15-17], the ALS data supports a $g_{OO}(r)$ exhibiting a taller and sharper first peak, and systematic shifts in all peak positions to smaller r (Figure 6). In what follows we evaluate the performance of various water models and simulation methodologies in reproducing this one property at this one thermodynamic state, with the reminder that this is where most models of water should perform quite well with the understanding that this is only one of many important properties of liquid water.

In the non-polarizable water model category we find that TIP3P[18], SPC[19], ST2[21], ST4[22], as well as MCY (data not shown) are inadequate structural descriptions of ambient water, while SPC/E[20] and TIP4P[18] give good agreement with our ALS experiment. The recently introduced TIP5P five-site model[23] gives excellent agreement with our ALS data (Figure 7), and with its robust performance for accurate densities over a large temperature range, makes it possibly the current non-polarizable water model of choice in classical simulation. We find that the NCC-vib[24], PPC[25], and TIP4P-FQ[26], and TIP4P-Pol-1[27] polarizable models perform well, but with some problems in the vicinity of the first peak. The CC model[28] has similar problems to the former polarizable models, but also shows shifts in all peak positions to larger *r* than what we determine from experiment.

In Figure 8 we show a comparison of our ALS-derived data with several recently reported *ab initio* simulated $g_{OO}(r)$'s. These include a 10ps, 64 water molecule MD run with a gradient-corrected BLYP functional with an average ionic temperature of 318K[29], a 5ps, 32 water molecule MD run with the gradient-corrected BLYP functional and average ionic temperature of 303K[30], a 2ps, 54 water molecule MD run with the PBE functional, and at a temperature of ~300K[31], and a new 6.4ps *ab initio* simulation with the PBE functional at ~294K by Schwegler and co-workers[32], to be reported in a future publication. The quantitative agreement is quite poor. The problems with the reported *ab initio* results arise from several sources that typically have been investigated and overcome in the classical simulation literature. These include dependence on initial conditions, length of the simulations, variation in system properties that arise with temperature or density, and finite size effects. Given the current computational expense of ab initio molecular dynamics that prohibit box sizes typically used in empirical force field simulations at present, these are largely technical limitations that will clearly diminish over time, and we would expect quantitative agreement to improve in the future.

Conclusions

The unification of our theoretical and experimental work is the development or discovery of effective protein interactions that implicitly includes the effects of aqueous solvent, and that potentially deeply influences the kinetics, thermodynamics, and their role in defining the intermediates on the folding pathway of proteins. While the experimental repertoire of protein-based structural techniques has resulted in a good understanding of a folding or folded protein's secondary structure and tertiary structure contacts, approaches to characterize the role of the hydration environment in terms of structure and forces in folding are comparatively minimal at present. We have combined our expertise in solution scattering experiments, simulations, and theory to address this deficiency. Furthermore, the benchmarking of protein simulations as to the quality of the underlying empirical force fields is naturally addressed in the simulation and experimental research proposed here. We hope that this work will lead forward to understanding and quantifying protein folding energy landscape biases and improvements in potential functions for water and proteins, but with an even longer term objective of taking protein-aqueous systems out of the context biology and towards design of a new biomaterials.

Acknowledgments. THG would like to acknowledge financial support from LDRD/DOE funds through U.S. Department of Energy Contract #DEAC-03-76SFOO098 and start-up funds from UC Berkeley. JMS thanks the National Science Foundation for a Graduate Research Fellowship. GH thanks the support of the NSF-IGERT training grant in Biophysics. We thank the ALS Macromolecular Crystallography Facility for use of the CCD area detector. We also thank Dr. Alastair McDowell for setting up the experimental station and Dr. Rich Celestre for assistance during the experimental runs at the ALS. We thank Dr. Bing Jap for use of his rotating anode source x-ray machine.

References

1. Onuchic, J., Luthey-Schulten, Z., & Wolynes, P. *Annu. Rev. Phys. Chem.* **48**, 545-600 (1997).
2. Baldwin R L. J. *Biomol. NMR* **5**, 103 (1995).
3. Shakhnovich, E. I. *Curr. Opin. Struct. Biol.* **7**, 29-40 (1997).
4. Z. Y. Guo & D. Thirumalai *Biopolymers* **36**, 83-102 (1994).
5. H. Nymeyer, A. E. Garcia, J. N. Onuchic. *Proc. Natl. Acad. Sci.* **95**, 5921 (1998).
6. J. M. Sorenson & T. Head-Gordon. *Proteins: Structure, Function, Genetics*. **37**, 582-91 (1999).
7. Sorenson, J. M. & Head-Gordon, T. *J. Comp. Bio*. **7**, 469-481 (2001).
8. J. M. Sorenson & T. Head-Gordon. *Proteins: Structure, Function, Genetics, in press* (2001).
9. J. M. Sorenson, G. Hura, A. Pertsemlidis, R. M. Glaeser & T. Head-Gordon. *Feature Article for J. Phys. Chem. B,* **103** 5413-5426 (1999).
10. A. Pertsemlidis, A. K. Soper, J. M. Sorenson & T. Head-Gordon. *Proc. Natl. Acad. Sci.* **96**, 481-486 (1999).
11. G. Hura, J. M. Sorenson, R. M. Glaeser & T. Head-Gordon. Perspectives in Drug Discovery and Design,**17**, 97-118 (1999).
12. G. Hura, J. Sorenson, R.M. Glaeser & T. Head-Gordon. *J. Chem. Phys*. **113**, 9140-9148 (2000).
13. J. Sorenson, G. Hura, R.M. Glaeser & T. Head-Gordon. *J. Chem. Phys* **113**, 9149-9161 (2000)..
14. Silvia Crivelli, Elizabeth Eskow, Brett Bader, Vincent Lamberti, Richard Byrd & Robert Schnabel, Teresa Head-Gordon. Submitted to Biophys. J. (2001).
15. A. H. Narten and H. A. Levy, J. Chem. Phys. **55**, 2263 (1971).
16. A. K. Soper and M. G. Phillips, Chem. Phys. **107**, 47 (1986).
17. A. K. Soper, F. Bruni, and M. A. Ricci. J. Chem. Phys. **106**, 247 (1997).
18. W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, and M.L. Klein. J. Chem. Phys. **79**, 926 (1983).
19. H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, and J. Hermans. In Intermolecular Forces, B. Pullman, editor, (D. Reidel Publishing Company, Dordrecht, 1981) 331.
20. H.J.C. Berendsen, J. R. Grigera, and T.P. Straatsma, J. Phys. Chem. **91**, 6269 (1987).
21. A. Rahman and F. H. Stillinger, J. Chem. Phys. **55**, 3336 (1971).
22. T. Head-Gordon and F. H. Stillinger, J. Chem. Phys. **98**, 3313 (1993).
23. M. W. Mahoney and W. L. Jorgensen, J. Chem. Phys. **112**, 8910 (2000).
24. G. Corongiu and E. Clementi, J. Phys. Chem. **97**, 2030 (1992).
25. I. M. Svishchev, P. G. Kusalik, J. Wang, and R. J. Boyd, J. Chem. Phys. **105**, 4742 (1996).
26. Y.-P. Liu, K. Kim, B. J. Berne, R. A. Friesner, and S. W. Rick, J. Chem. Phys. **108**, 4739 (1998).
27. B. Chen, J. Xing and J. I. Siepmann, J. Phys. Chem. B **104**, 2391 (2000).
28. A. A. Chialvo and P. T. Cummings, J. Chem. Phys. **105**, 8274 (1996).
29. P. L. Silvestrelli and M. Parrinello, J. Chem. Phys. **111**, 3572 (1999).
30. M. Sprik, J. Hutter, and M. Parrinello, J. Chem. Phys. **105**, 1142 (1996).
31. E. Schwegler, G. Galli, and F. Gygi, Phys. Rev. Lett. **84**, 2429 (2000).
32. E. Schwegler and G. Galli (unpublished results).

Figure Captions

Figure 1. Experimental x-ray scattering intensity curves for pure water and NALMA in water. Mole ratios of solute to water of ~1:25 (black), ~1:50 (gray dots), and pure water (gray squares).

Figure 2. Comparison of the x-ray solution scattering experiment and three simulated intensity curves for three different $g_c(r)$ models for NALMA in SPC water at concentrations of solute to water of 1:24.

Figure 3. The potential of mean force consistent with the aqueous models of $g_c(r)$ that reproduce the leucine-leucine correlations over a range of concentrations.

Figure 4. Using the TIP-FQ water with AMBER parameterization of NALMA shows better quantitative agreement with experiment.

Figure 5. Comparison of current experimental $g_{OO}(r)$ with previous work. The fit was obtained with α=1.333 and δ=2.2 $Å^{-1}$. Legend: Narten and Levy, xray (grey line); Soper, Bruni, and Ricci, neutron (dash line); ALS data, x-ray (black line).

Figure 6. A comparison of x-ray scattering experimental data on pure water at 25C and 1atm as determined by our recent synchrotron ALS work (black), that of Narten (gray), and of Nishikawa (light gray).

Figure 7. Comparison of ALS experimental $g_{OO}(r)$ (black line) with the newest water model TIP5P (gray line) that shows the best agreement compared to all simulations of liquid water examined in this paper.

Figure 8. Comparison of ALS experimental $g_{OO}(r)$ (black line) with ab initio molecular dynamics simulations. Carr-Parinello molecular dynamics (CPMD) run of 10ps for 64 water molecules, average ionic temperature of 318°K (dotted line), CPMD run of 12ps for 32 water molecules, average ionic temperature of 303°K[60] (dot-dash line), CPMD run of 2ps for 54 water molecule average ionic temperature of ~300°K[61] (grey).

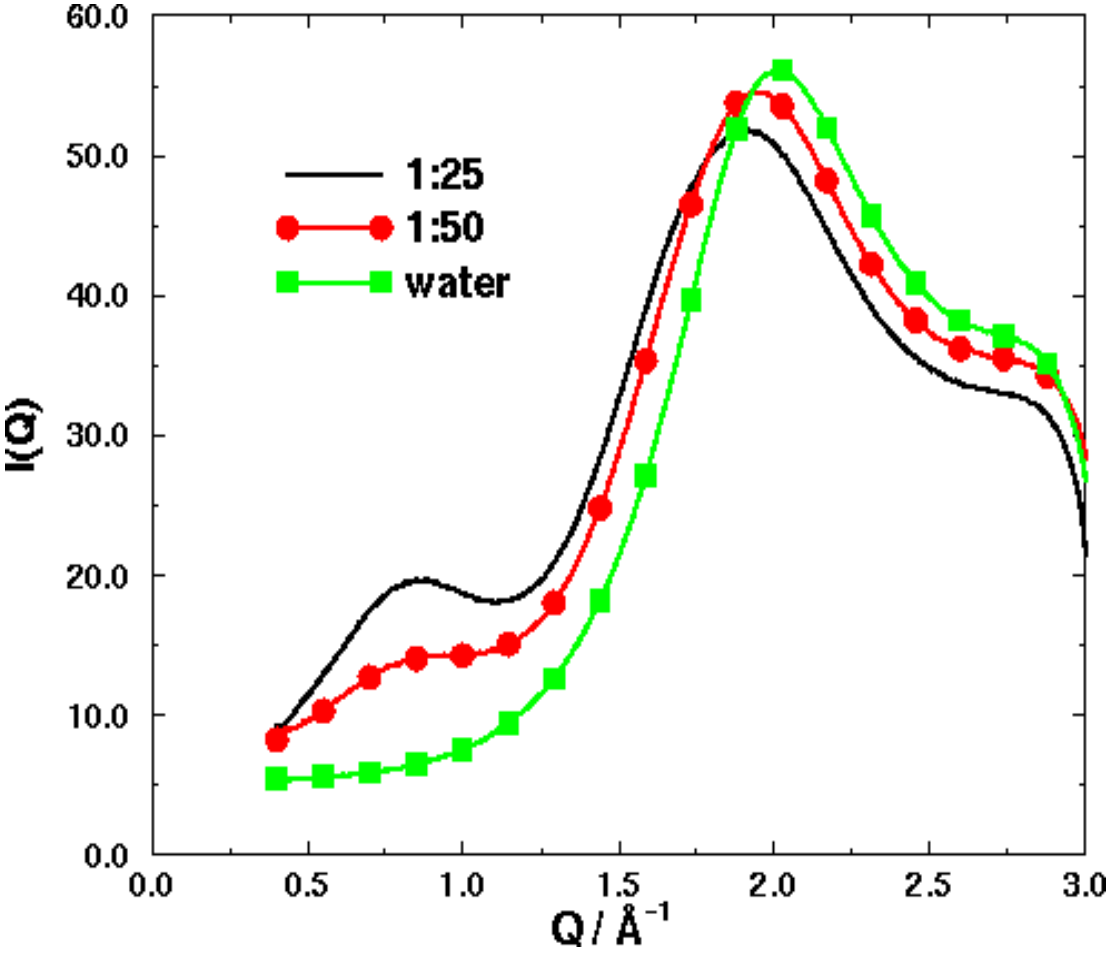

Figure 1. Head-Gordon

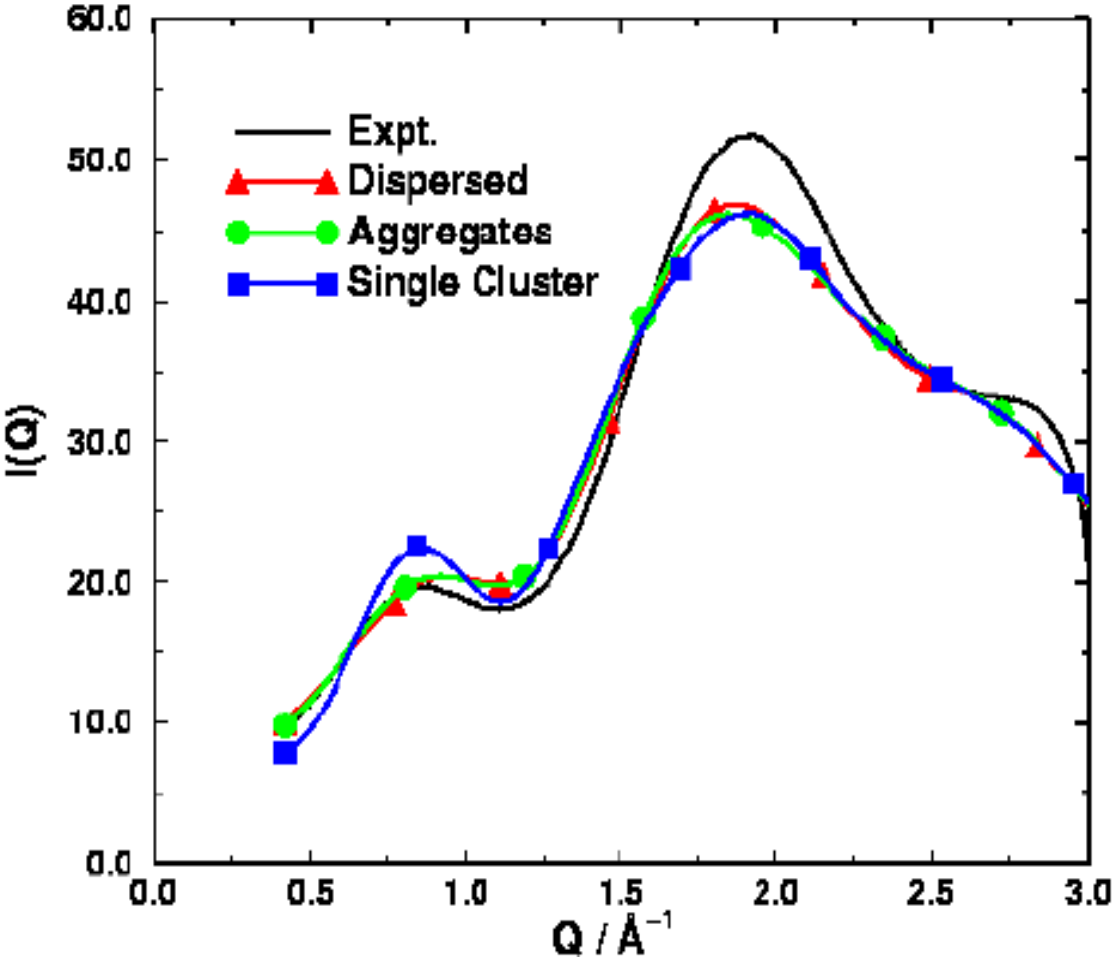

Figure 2. Head-Gordon

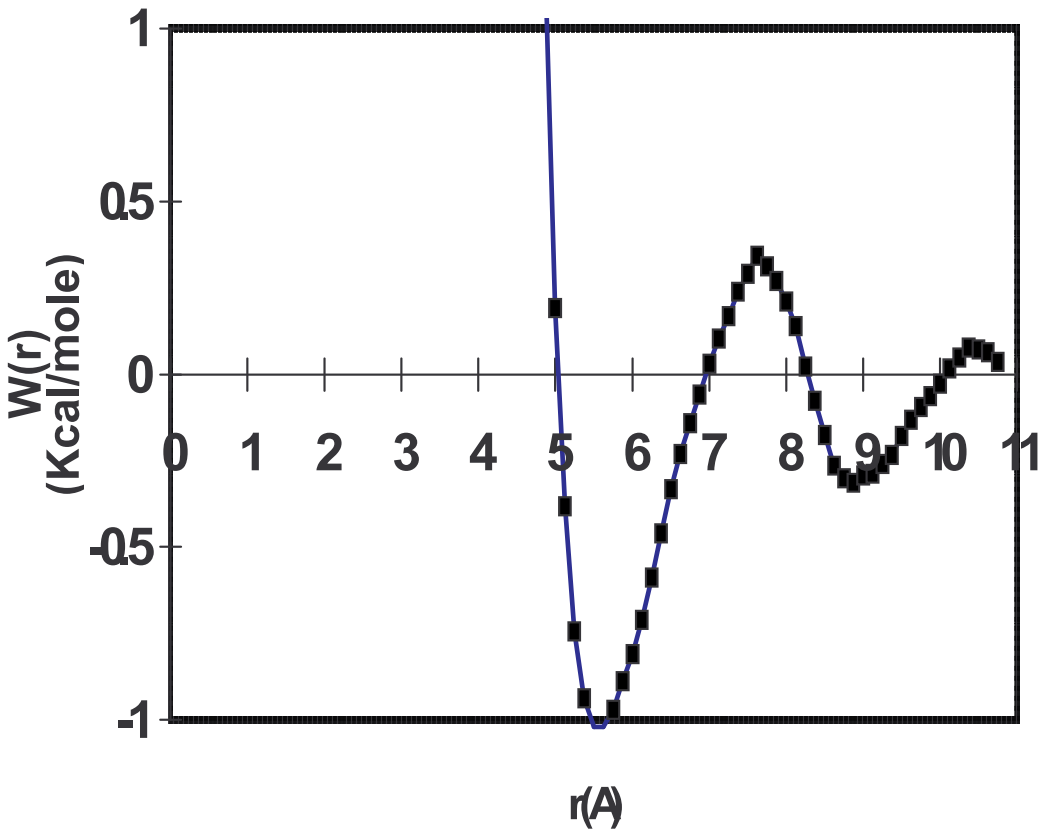

Figure 3. Head-Gordon

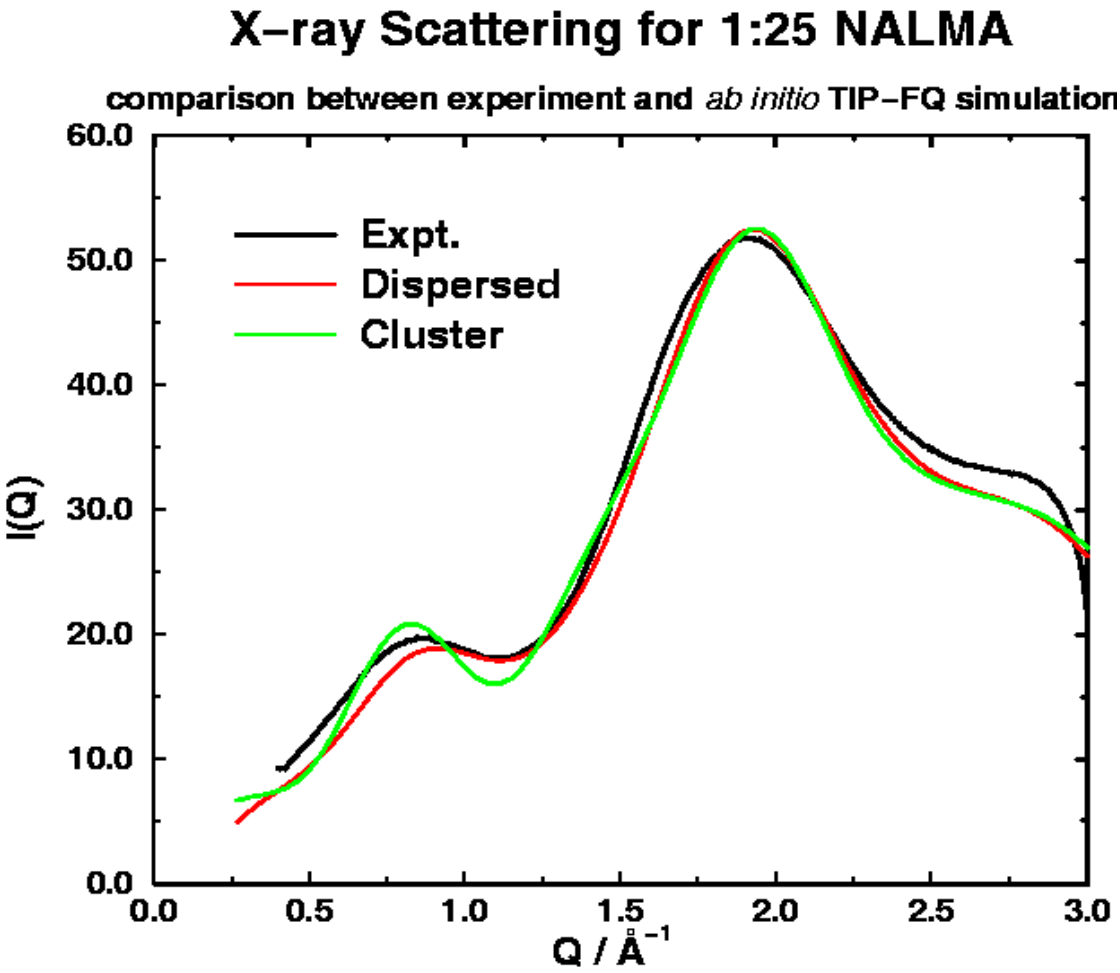

Figure 4. Head-Gordon

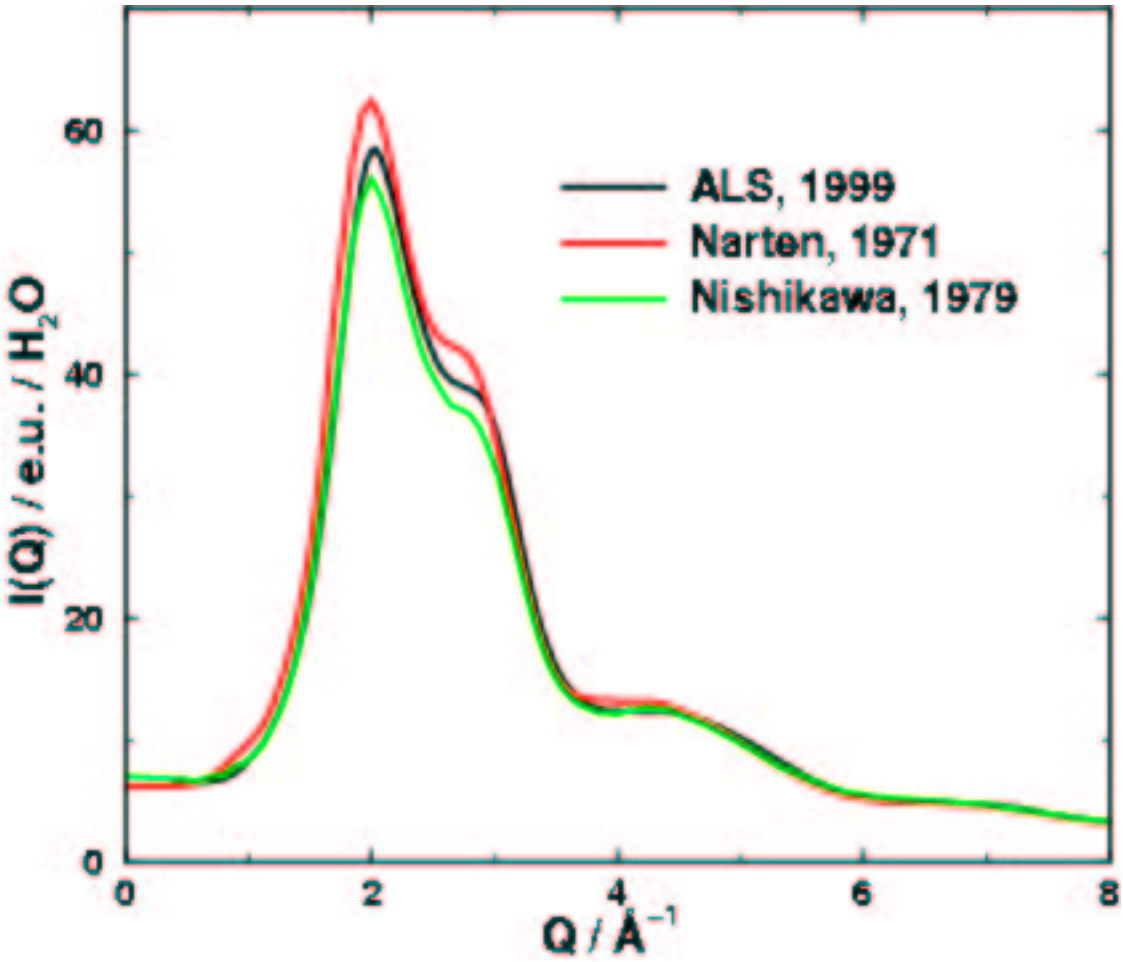

Figure 5. Head-Gordon

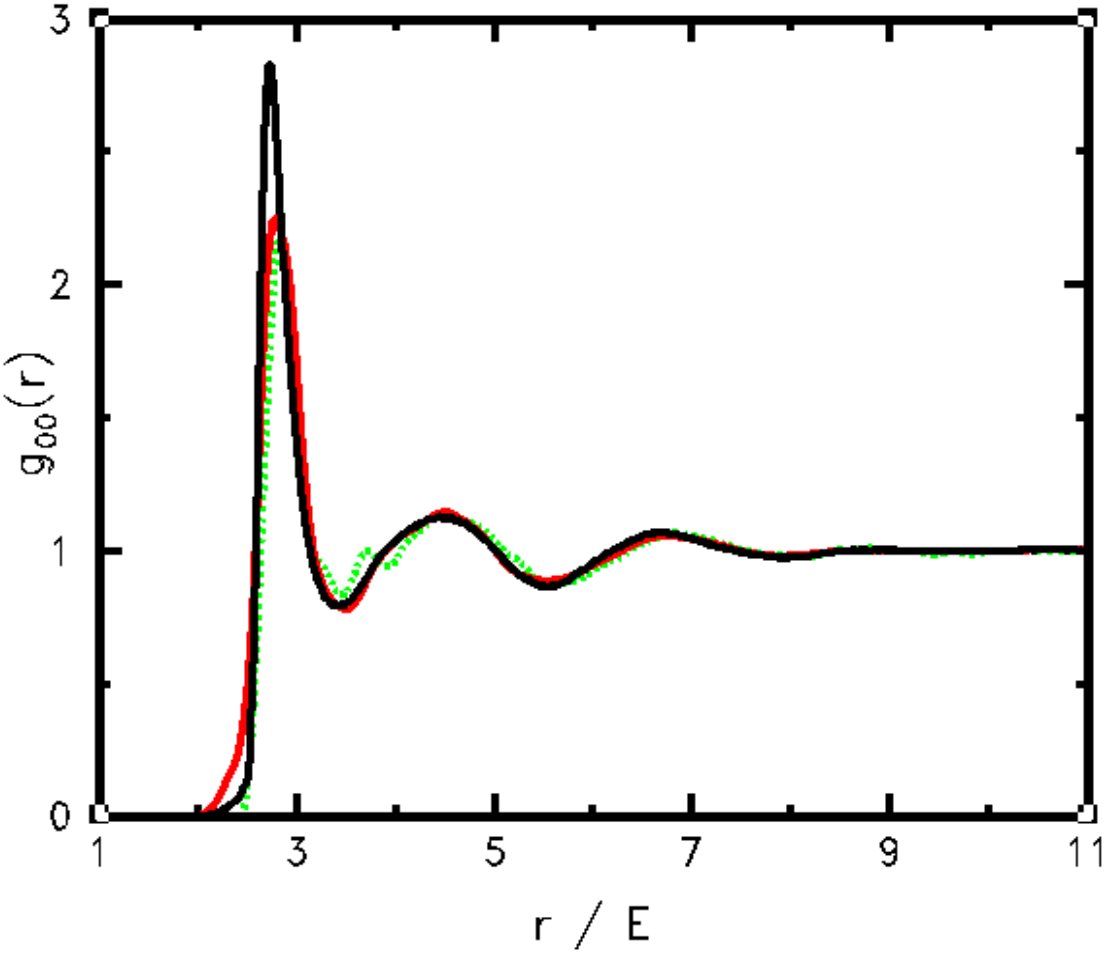

Figure 6. Head-Gordon

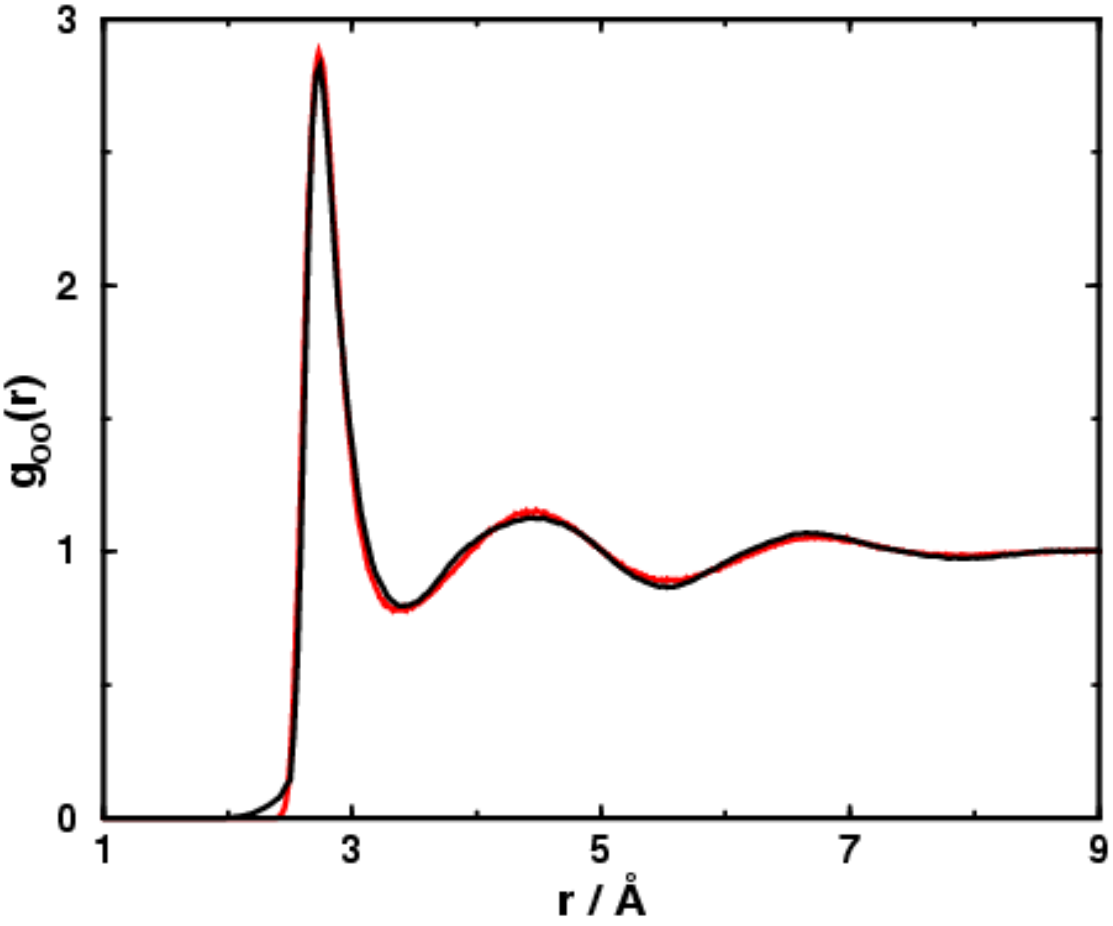

Figure 7. Head-Gordon

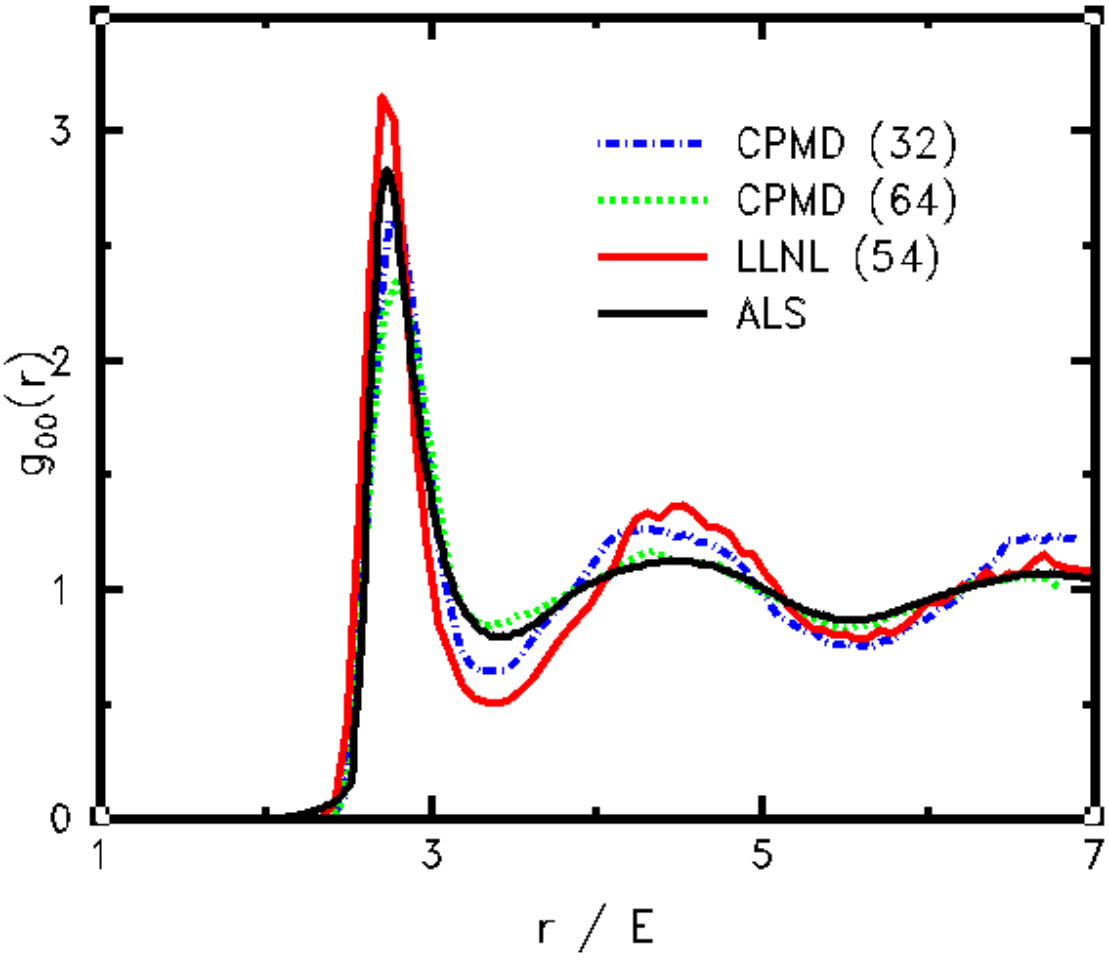

Figure 8. Head-Gordon